Product Requirements Document (PRD)

Project Title: MPEG-G Decoding the Dialogue – Track 1: Predicting Cytokine Profiles from Metagenomic Data

Team:

Desmond Brown — Level 100 Computer Engineering student, Professional Software Engineer with strong Python background, beginner in ML & Neural Networks, actively learning. Responsible for code structure, data pipelines, advanced model experimentation, and learning both ML and the biological domain (metagenomics, cytokines, microbiomes).
Sarah Amewu — Level 100 Biomedical Engineering student, beginner in ML and metagenomics, responsible for domain research, basic data preparation, assisting in experimentation, and also learning ML, metagenomics, cytokines, and microbiomes.

1. Objective

To develop and submit a predictive model for Track 1 of the competition that, given metagenomic data, predicts cytokine profiles for samples. Secondary objective: use this competition as a practical learning pathway to master ML techniques, deepen understanding of neural networks, and gain foundational knowledge in metagenomics, microbiome science, and host–microbiome interactions.

2. Background

The competition focuses on multi-output regression: mapping high-dimensional microbiome-related features (possibly taxonomic and/or functional) to multiple cytokine concentration values. Challenges include processing metagenomic data (possibly in MPEG-G format), building preprocessing pipelines, handling multi-output learning with correlated outputs, and applying appropriate validation schemes. Both team members are new to metagenomics and will learn the biological and computational aspects together.

3. Goals & Success Criteria

Primary Success Criteria:
- Submit at least one valid model prediction to Zindi.
- Achieve better than baseline performance provided by organizers.

Secondary Success Criteria:
- Implement at least three different model families (linear, tree-based, neural net).
- Create a reproducible data pipeline.
- Document modeling decisions, preprocessing steps, and lessons learned.
- Both team members can explain multi-output regression, feature engineering for compositional data, and basic metagenomics concepts.

4. Scope

In Scope:
- Understanding dataset structure.
- Parsing MPEG-G or pre-processed profiles.
- Data cleaning, transformation, and feature engineering.
- Implementation of multiple model types.
- Grouped cross-validation.
- Ensembling models.
- Submitting predictions in required format.

Out of Scope:
- Full bioinformatics pipelines beyond minimal decoding.

- Extensive hyperparameter tuning without available resources.

5. Functional Requirements

Data Pipeline:
- Input: Training features, targets, and test features.
- Processing: Decode MPEG-G if needed, feature engineering (CLR, diversity indices, optional PCA).
- Output: Feature matrix X, Target matrix Y.

Modeling:
- Train multiple model families.
- Grouped cross-validation.
- Ensemble best models.

Evaluation:
- Match competition's scoring metric.
- Report per-cytokine and average performance.

Submission:
- Predictions match required format.
- Ensure reproducibility.

6. Non-Functional Requirements

- Reproducibility with fixed seeds.
- Collaboration via GitHub.
- Documentation via README.
- Learning focus: document both ML and metagenomics insights.

7. Milestones & Timeline

Week 1	Understand rules, set up repo, inspect data	Desmond, Sarah
Week 1	Research MPEG-G, cytokines, microbiome basics	Both
Week 1–2	Implement data parsing + baseline linear model	Desmond
Week 2	Implement LightGBM/XGBoost models	Desmond
Week 3	Implement simple MLP	Desmond
Week 3	Document learning so far	Both
Week 4	Hyperparameter tuning & ensembling	Both
Week 4	Final submission prep & write-up	Both

8. Risks & Mitigation

- Unfamiliarity with metagenomics: both members research together.
- Overfitting: grouped CV and regularization.
- MPEG-G decoding complexity: confirm pre-decoded availability early.
- Balancing learning vs delivery: split tasks strategically.

9. Learning Outcomes

By project end, both Desmond and Sarah will:
- Understand end-to-end ML pipeline.
- Have practical multi-output regression experience.

- Gain foundational metagenomics, microbiome, and cytokine knowledge.
- Connect computational predictions with biological meaning.